Pagani, L., Lawson, D., Jagoda, E., Morseburg, A., Eriksson, A., Mitt, M., Kivisild, T., & Metspalu, M. (2016). Genomic analyses inform on migration events during the peopling of Eurasia. *Nature*, *538*(7624), 238-242. https://doi.org/10.1038/nature19792

Environmental challenges and complex migration events during the peopling of Eurasia

Authors List

Luca Pagani[1,2*†], Daniel John Lawson[3*], Evelyn Jagoda[1,4*], Alexander Mörseburg[1*], Anders Eriksson[5,6*],Mario Mitt[7,8], Florian Clemente[1,9], Georgi Hudjashov[10,11,12], Michael DeGiorgio[13], Lauri Saag[10], Jeffrey D. Wall[14], Alexia Cardona[1,15], Reedik Mägi[7], Melissa A. Wilson Sayres[16,17], Sarah Kaewert[1], Charlotte Inchley[1], Christiana L. Scheib[1], Mari Järve[10], Monika Karmin[10,18], Guy S. Jacobs[19,20], Tiago Antao[21], Florin Mircea Iliescu[1], Alena Kushniarevich[10,22], Qasim Ayub[23], Chris Tyler-Smith[23], Yali Xue[23], Bayazit Yunusbayev[10,24], Kristiina Tambets[10], Chandana Basu Mallick[10], Lehti Saag[18], Elvira Pocheshkhova[25], George Andriadze[26], Craig Muller[27], Michael C. Westaway[28], David M. Lambert[28], Grigor Zoraqi[29], Shahlo Turdikulova[30], Dilbar Dalimova[31], Zhaxylyk Sabitov[32], Gazi Nurun Nahar Sultana[33], Joseph Lachance[34,35], Sarah Tishkoff[36], Kuvat Momynaliev[37], Jainagul Isakova[38], Larisa D. Damba[39], Marina Gubina[39], Pagbajabyn Nymadawa[40], Irina Evseeva[41,42], Lubov Atramentova[43], Olga Utevska[43], François-Xavier Ricaut[44], Nicolas Brucato[44], Herawati Sudoyo[45], Thierry Letellier[44], Murray P. Cox[12], Nikolay A. Barashkov[46,47], Vedrana Skaro[48,49], Lejla Mulahasanovic[50], Dragan Primorac[51,52,53,49], Hovhannes Sahakyan[10,54], Maru Mormina[55], Christina A. Eichstaedt[1,56], Daria V. Lichman[39,57], Syafiq Abdullah[58], Gyaneshwer Chaubey[10], Joseph T. S. Wee [59], Evelin Mihailov[7], Alexandra Karunas[24,60], Sergei Litvinov[24,60,10], Rita Khusainova[24,60], Natalya Ekomasova[60], Vita Akhmetova[24], Irina Khidiyatova[24,60], Damir Marjanovic´[61,62], Levon Yepiskoposyan[54], Doron M. Behar[10], Elena Balanovska[63], Andres Metspalu[7,8], Miroslava Derenko[64], Boris Malyarchuk[64], Mikhail Voevoda[65,39,57], Sardana A. Fedorova[47,46], Ludmila P. Osipova[39,57], Marta Mirazón Lahr[66], Pascale Gerbault[67], Matthew Leavesley[68,69], Andrea Bamberg Migliano[70], Michael Petraglia[71], Oleg Balanovsky[72,63], Elza K. Khusnutdinova[24,60], Ene Metspalu[10,18], Mark G. Thomas[67], Andrea Manica[6], Rasmus Nielsen[73], Richard Villems[10,18,74*], Eske Willerslev[27*], Toomas Kivisild[1,10*†], Mait Metspalu[10,18*†]

[*]*These authors contributed equally to this work.*

†*Corresponding authors: L.P. (lp.lucapagani@gmail.com), T.K. (tk331@cam.ac.uk), M.M. (mait@ebc.ee)*

Author Affiliations

1: Department of Biological Anthropology, University of Cambridge, Cambridge, United Kingdom

2: Department of Biological, Geological and Environmental Sciences, University of Bologna, Via Selmi 3, 40126, Bologna, Italy

3: Integrative Epidemiology Unit, School of Social and Community Medicine, University of Bristol, Bristol BS8 2BN, UK.

4: Department of Human Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA

5: Integrative Systems Biology Lab, Division of Biological and Environmental Sciences & Engineering, King Abdullah University of Science and Technology, Thuwal, Kingdom of Saudi Arabia

6: Department of Zoology, University of Cambridge, Cambridge, UK

7: Estonian Genome Center, University of Tartu, Tartu, Estonia

8: Department of Biotechnology, Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia

9: Institut de Biologie Computationnelle, Université Montpellier 2, Montpellier, France

10: Estonian Biocentre, Tartu, Estonia

11: Department of Psychology, University of Auckland, Auckland, 1142, New Zealand;

12: Statistics and Bioinformatics Group, Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand

13: Department of Biology, Pennsylvania State University, University Park, PA, 16802, USA

14: Institute for Human Genetics, University of California, San Francisco, California 94143, USA

15: MRC Epidemiology Unit, University of Cambridge, Institute of Metabolic Science, Box 285, Addenbrooke's Hospital, Hills Road, Cambridge, CB2 0QQ

16: School of Life Sciences, Tempe, AZ, 85287 USA

17: Center for Evolution and Medicine, The Biodesign Institute, Tempe, AZ, 85287 USA

18: Department of Evolutionary Biology, Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia

19: Mathematical Sciences, University of Southampton, Southampton SO17 1BJ, UK

20: Institute for Complex Systems Simulation, University of Southampton, Southampton SO17 1BJ, UK

21: Division of Biological Sciences, University of Montana, Missoula, MT, USA

22: Institute of Genetics and Cytology, National Academy of Sciences, Minsk, Belarus

23: The Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, United Kingdom

24: Institute of Biochemistry and Genetics, Ufa Scientific Center of RAS, Ufa , Russia

25: Kuban State Medical University, Krasnodar, Russia

26: Scientific-Research Center of the Caucasian Ethnic Groups, St. Andrews Georgian University, Georgia

27: Center for GeoGenetics, University of Copenhagen, Denmark

28: Research Centre for Human Evolution, Environmental Futures Research Institute, Griffith University, Nathan, Australia

29: Center of Molecular Diagnosis and Genetic Research, University Hospital of Obstetrics and Gynecology, Tirana, Albania

30: Center of High Technology, Academy of Sciences, Republic of Uzbekistan

31: Institute of Bioorganic Chemistry Academy of Science, Republic of Uzbekistan

32: L.N. Gumilyov Eurasian National University, Astana, Kazakhstan

33: Centre for Advanced Research in Sciences (CARS), DNA Sequencing Research Laboratory, University of Dhaka, Dhaka-1000, Bangladesh

34: Department of Genetics, University of Pennsylvania, Philadelphia, PA, 19104-6145, USA

35: School of Biology, Georgia Institute of Technology, Atlanta, Georgia, USA

36: Departments of Genetics and Biology, University of Pennsylvania, Philadelphia, Pennsylvania, USA

37: DNcode laboratories, Moscow, Russia

38: Institute of Molecular Biology and Medicine, Bishkek, Kyrgyz Republic

39: Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia

40: Mongolian Academy of Medical Sciences, Ulaanbaatar, Mongolia

41: Northern State Medical University, Arkhangelsk, Russia

42: Anthony Nolan, London, United Kingdom

43: V. N. Karazin Kharkiv National University, Kharkiv, Ukraine

44: Evolutionary Medicine group, Laboratoire d'Anthropologie Moléculaire et Imagerie de Synthèse, UMR 5288, Centre National de la Recherche Scientifique, Université de Toulouse 3, Toulouse, France

45: Genome Diversity and Diseases Laboratory, Eijkman Institute for Molecular Biology, Jakarta, Indonesia

46: Department of Molecular Genetics, Yakut Scientific Centre of Complex Medical Problems, Yakutsk, Russia

47: Laboratory of Molecular Biology, Institute of Natural Sciences, M.K. Ammosov North-Eastern Federal University, Yakutsk, Russia

48: Genos, DNA laboratory, Zagreb, Croatia

49: University of Osijek, Medical School, Osijek, Croatia

50: Center for Genomics and Transcriptomics, CeGaT, GmbH, Tübingen, Germany

51: St. Catherine Speciality Hospital, Zabok, Croatia

52: Eberly College of Science, The Pennsylvania State University, University Park, PA, USA

53: University of Split, Medical School, Split, Croatia

54: Laboratory of Ethnogenomics, Institute of Molecular Biology, National Academy of Sciences, Republic of Armenia, 7 Hasratyan Street, 0014, Yerevan, Armenia

55: Department of Applied Social Sciences, University of Winchester, Sparkford Road, Winchester SO22 4NR, UK

56: Thoraxclinic at the University Hospital Heidelberg, Heidelberg, Germany

57: Novosibirsk State University, Novosibirsk, Russia.

58: RIPAS Hospital, Bandar Seri Begawan, Brunei Darussalam

59: National Cancer Centre Singapore, Singapore

60: Department of Genetics and Fundamental Medicine, Bashkir State University, Ufa , Russia

61: Department of Genetics and Bioengineering. Faculty of Engineering and Information Technologies, International Burch University, Sarajevo, Bosnia and Herzegovina

62: Institute for Anthropological Researches, Zagreb, Croatia

63: Research Centre for Medical Genetics, Russian Academy of Sciences, Moscow 115478, Russia

64: Genetics Laboratory, Institute of Biological Problems of the North, Russian Academy of Sciences, Magadan, Russia

65: Institute of Internal Medicine, Siberian Branch of Russian Academy of Medical Sciences, Novosibirsk, Russia

66: Leverhulme Centre for Human Evolutionary Studies, Department of Archaeology and Anthropology, University of Cambridge, Cambridge, United Kingdom

67: Research Department of Genetics, Evolution and Environment, University College London, London, United Kingdom

68: Department of Archaeology, University of Papua New Guinea, University PO Box 320, NCD, Papua New Guinea

69: College of Arts, Society and Education, James Cook University, PO Box 6811, Cairns QLD 4870, Australia

70: Department of Anthropology, University College London, London, United Kingdom

71: Max Planck Institute for the Science of Human History, Kahlaische Strasse 10, D-07743 Jena, Germany

72: Vavilov Institute for General Genetics, Russian Academy of Sciences, Moscow, Russia

73: Department of Integrative Biology, University of California Berkeley, Berkeley 94720, CA, USA

74: Estonian Academy of Sciences, 6 Kohtu Street, Tallinn 10130, Estonia

High-coverage whole-genome sequence studies have so far focused on a small number[1] of geographically restricted populations [2-5], or targeted at specific diseases, e.g. cancer[6]. Nevertheless, the availability of high-resolution genomic data has led to the development of new methodologies for inferring population history[7-9] and refuelled the debate on the mutation rate in humans[10].

Here we present the Estonian Biocentre human Genome Diversity Panel (EGDP), a dataset of 483 high-coverage human genomes from 148 populations worldwide, including 379 new genomes from 125 populations, which we group into Diversity and Selection Sets (ED1-2; SI1:1.1-7). The combination of high spatial and genomic coverage enabled us to refine current knowledge of continent-wide patterns of heterozygosity, long- and short-distance gene flow, archaic admixture, and changes in effective population size through time. Our most surprising find is a genetic signature in present-day Papuans consistent with an early and largely extinct expansion of anatomically modern humans (AMH) Out-of-Africa (xOoA). Modelling shows that this genetic signature may represent an early and largely extinct expansion of modern humans Out-of-Africa (xOoA), seen in the Western Asian fossil record[11], and is consistent with admixture between AMHs and Neanderthals predating the main Eurasian expansion [12]. We also identify number of new metabolism- and immunity-related loci as candidates for local adaptation based on signals of positive or balancing selection.

The paths taken by AMHs out of Africa (OoA) has been the subject of considerable debate over the past two decades. Fossil and archaeological evidence[13,14], and craniometric studies[15] of African and Asian populations, demonstrate that *Homo sapiens* was present outside of Africa ca. 120-70 kya[11]. However, this colonization has been viewed as a failed expansion OoA[16] since genetic analyses of living populations have been consistent with a single OoA followed by serial founder events[17].

Ancient DNA (aDNA) sequencing has revealed admixture between early Eurasians and at least two archaic human lineages[18,19], and suggests modern human reached Eurasia at around 100kya[12]. In addition, aDNA from modern

humans suggests population structuring and turnover, but little additional archaic admixture, in Eurasia over the last 35-45 thousand years[20-22]. Overall, these findings indicate that the majority of human genetic diversity outside Africa derives from a single dispersal event that was followed by admixture with archaic humans[18,23].

We used ADMIXTURE to visualise the genetic structure in our Diversity Set (ED1). We further compared the individual-level haplotype similarity of our samples using fineSTRUCTURE (ED3). Despite small sample sizes, we inferred 106 genetically distinct populations forming 12 major regional clusters, corresponding well to the 148 self-identified population labels. This clustering forms the basis for the groupings used in the scans of natural selection. Similar genetic affinities are highlighted by plotting the outgroup f_3 statistic[9] in the form $f_3(X, Y;$ Yoruba), which here measures shared drift between non-African populations X modern and aDNA Y from Yoruba as an African outgroup (SI1:2.2.6, ED4).

Our geographically dense sampling allowed us to quantify global barriers for gene flow by spatially interpolating genetic similarity measures between pairs of populations (SI1:2.2.2). We considered several measures and report gradients of allele frequencies in Figure 1 (validated by comparing to gene flow patterns from EEMS[24], ED5). Controlling for pairwise geographic distance, we show that the genetic gradients are linked to geographic and climatic features, most importantly precipitation and elevation (inset of Figure 1, SI1:2.2.2).

In addition to these geographical barriers, humans also faced a number of new ecological challenges as they expanded out of Africa. To identify potential resultant adaptations, we explored the distribution of functional variants among populations, performed tests of purifying, balancing and positive selection and, finally, identified loci that showed the highest allelic differentiation among groups (SI1:3). Our positive and purifying selection scans (Methods) corroborated some previously known and functionally-supported findings (SI2:3.3.4-I, SI1:3.1, ED6; SI2:3.1-IV,VI). Additionally, we infer more purifying selection in Africans in genes involved in pigmentation (bootstrapping p value - bpv for $R_{X/Y}$-scores <0.05) (ED6) and immune response against viruses (bpv<0.05), whilst further purifying selection was indicated on olfactory

receptor genes in Asians (bpv $p<0.05$) (SI2:3.1.1-II). Our scans for ancient balancing selection found a significant enrichment (FDR <0.01) of antigen processing/presentation, antigen binding, and MHC and membrane component genes (SI1:3.2, SI2:3.3.2-I-III). The HLA (*HLA-C*)-associated gene (*BTNL2*) was the top candidate in eight of 12 geographic regions (SI2:3.3.1-I).

Our positive selection scans, variant-based analyses (SI1:3.2 and 3.2) and gene enrichment studies revealed many novel signals (SI1:3.4; SI2:3.5-I-VI), a subset of which is highlighted in Table 1. We were also able to identify new potentially causal variants in novel and previously-detected signals (SI1:3).

The largest demographic outlier in our Diversity Set is an excess of short African haplotypes in Papuans, as well as Philippine Negritos, compared to all other non-African populations (ED7). This pattern remains after correcting for potential confounders such as phasing errors and sampling bias (SI1:2.2.1). These shorter shared haplotypes would be consistent with an older population split[25]. Indeed, the Papuan-Yoruban median genetic split time (using MSMC) of 90 kya predates the split of all mainland Eurasian populations from Yorubans at ~75 kya (SI2:2.2.3-I, ED4, Figure 2A). This result is robust to phasing artefacts (ED8, See Methods). Furthermore, the Papuan-Eurasian MSMC split time of ~40 kya is older than splits between West Eurasian and East Asian populations at ~30 kya (ED4). The Papuan split times from Yoruba and Eurasia are therefore incompatible with a simple bifurcating population tree model. Finally, in large sections of the Papuan genome, Papuans form an outgroup to most modern Africans and Eurasians, while the rest of their genome forms a clade with Eurasia.

At least two main models could account for Sahul populations having older split dates from Africa than mainland Eurasians in our sample: 1) Admixture in Sahul with a potentially un-sampled archaic human population that split from modern humans either before or at the same time as did Denisova and Neanderthal; or 2) Admixture in Sahul with a modern human population (xOoA) that left Africa after the split between modern humans and Neanderthals, but before the main expansion of modern humans in Eurasia (main OoA).

We performed extensive analysis to identify the source of the strongest contribution from these two non-mutually exclusive scenarios. Because the

introgressing lineage has not been observed with aDNA, standard methods are limited in their ability to distinguish between these hypotheses. Furthermore we show (SI1:2.2.7) that single-site statistics, such as Patterson's *D* [9,18] and sharing of non-African Alleles (nAAs), are inherently affected by confounding effects due to archaic introgression in non-African populations[23]. Our approach therefore relies on building multiple lines of evidence using haplotype-based MSMC and fineSTRUCTURE comparisons (which we show should have power at this time scale[26]; SI2.2.13).

We located and masked putatively introgressed[27] Denisova haplotypes from the genomes of Papuans, and evaluated phasing errors by symmetrically phasing Papuans and Eurasians (Methods). Neither modification (Figure 3A, SI1:2.2.9, SI2:2.2.9-I) changed the estimated split time (based on MSMC) between Africans and Papuans (Methods, SI1:2.2.8, ED8, Table 2.2.8-I). MSMC dates behave approximately linearly under admixture (ED8), implying that the hypothesised lineage may have split from most Africans around 120 kya (SI1:2.2.4 and 2.2.8).

We compared the effect on the MSMC split times of a xOoA or a Denisova lineage in Papuans by extensive coalescent simulations (SI1:2.2.8). We could not simulate the large Papuan-African and Papuan-Eurasian split times inferred from the data, unless assuming an implausibly large contribution from a Denisova-like population. Furthermore, while the observed shift in the African-Papuan MSMC split curve can be qualitatively reproduced when including a 4% genomic component that diverged 120 kya from the main human lineage within Papuans, a similar quantity of Denisova admixture does not produce any significant effect (ED8). This favours a small presence of xOoA lineages rather than Denisova admixture alone as the likely cause of the observed deep African-Papuan split. We also show (Methods) that such a scenario is compatible with the observed mtDNA and Y chromosome lineages in Oceania, as also previously argued[13,28].

We further tested our hypothesised xOoA model by focussing on genomic regions in Papuans that have African ancestry not found in other Eurasian populations. We re-ran fineSTRUCTURE adding the Denisova, Altai Neanderthal and the Human Ancestral Genome sequences[29] to a subset of the Diversity Set. FineSTRUCTURE infers chunks of the genome that have a most recent common

ancestor (MRCA) with another individual. Papuan chunks assigned African had, regardless, an elevated level of non-African derived alleles (i.e. nAAs fixed ancestral in Africans) compared to such chunks in Eurasians. They therefore have an older mean coalescence time with our African samples.

Due to the deep divergence between the sampled Denisova and the one introgressed into modern humans, it is possible that some archaic haplotypes have a MRCA with an African instead of Denisova and are assigned as "African". We can resolve the coalescence time, and hence origin, of these chunks by their sequence similarity with modern Africans. To account for the archaic introgression we modelled these genomic segments as a mixture of chunks assigned a) African or b) Denisova in Eurasians and c) chunks assigned Denisova in Papuans. Chunks are modelled (see Methods, ED9) in terms of the distribution of length and mutation rate measured as a density of non-African derived alleles. Since Eurasians (specifically Europeans) have not experienced Denisova admixture, this approach disentangles lineages that coalesce before the human/Denisova split from those that coalesce after.

We found that the xOoA component (Figure 2B-D; SI1:2.2.10) was necessary to account for the number of short chunks with "moderate" nAAs density in the data (i.e. proportion of non-African derived sites higher than that of Eurasian chunks assigned African but significantly lower than that of those assigned Denisova in either Eurasians or Papuans). Consistent with our MSMC findings (SI1:2.2.4), xOoA chunks have an estimated MRCA 1.5 times older than the Eurasian chunks in Papuan genomes, while the Denisovan chunks in Papuans are 4 times older than the Eurasian chunks. Adding up the contributions across the genome (Methods) leads to a genome-wide estimate of 1.9% xOoA (95% CI 1.5-3.3) in Papuans, which we view as a lower bound.

Our results consistently point towards a contribution from a modern human source for derived[29] alleles that are found in Papuans but not in Africans. Possible confounders could involve a shorter generation time in Papuan and Philippine Negrito populations[30], different recombination processes, or alternative demographic histories that have not been investigated here. We therefore strongly encourage the development of new model-based approaches that can investigate further the haplotype patterns described here.

In conclusion, our results suggest that while the genomes of modern Papuans derive primarily from the main expansion of modern humans out of Africa, at least 2% of their genome is retained from an earlier, otherwise extinct, dispersal (ED10).

The inferred date of the xOoA split time (~120 kya) is consistent with fossil and archaeological evidence for an early expansion of *Homo sapiens* from Africa[13,14]. Furthermore, the recently identified modern human admixture into the Altai Neanderthal before 100 kya[12] is consistent with a modern human presence outside Africa well before the main OoA split time (~75 kya). Further studies will confirm whether the Papuan genetic signature reported here and the one observed in Altai Neanderthals were carved by the same xOoA human group, as well as clarify the timing and route followed during such an early expansion. The high similarity between Papuans and the Altai Neanderthal reported in ED1 may indeed reflect a shared xOoA component. The unexpected genetic traces of xOoA in Papuans, for which we show evidence here for the first time, suggest that unravelling the evolutionary history of our own species will require the recovery of aDNA from additional fossils, and further archaeological investigations in under-explored geographical regions.

Data availability

The newly sequenced genomes are part of the Estonian Biocentre human Genome Diversity Panel (EGDP) and were deposited in the ENA archive under accession number PRJEB12437 and are also freely available through the Estonian Biocentre website (www.ebc.ee/free_data).

Supplementary Information:

Additional results are reported in two Supplementary Information files online:
SI1 including description of additional analyses, and SI2 including results in table format.

Author Contributions:

Conceived the study: R.V, E.W, T.K, M.M.

Conducted anthropological research and/or sample collection and management: A.K, K.T, C.B.M, Le.S, E.P, G.A, C.M, M.W, D.L, G.Z, S.T, D.D, Z.S, G.N.N.S, K.M, J.I, L.D.D, M.G, P.N, I.E, L.At, O.U, F.-X.R, N.B, H.S, T.L, M.P.C, N.A.B, V.S, L.A, D.Pr, H.Sa, M.Mo, C.A.E, D.V.L, S.A, G.C, J.T.S.W, E.Mi, A.Ka, S.L, R.K, N.T, V.A, I.K, D.M, L.Y, D.M.B, E.B, A.Me, M.D, B.M, M., S.A.F, L.P.O, M.M., M.L, A.B.M, O.B, E.K.K, E.M, M.G.T, E.W.

Provided access to data: J.L, S.Ti.

Analysed data: L.P, D.J., E.J, A.M, M.Mit, F.C, G.H, M.D, A.E, L.S, J., A.C, R.M, M.A.W.S, S.K, C.I, C.L.S, M.J, M.K, G.S.J, T., F.M.I, A.K, Q.A, C.T.-S, Y.X, B.Y, C.B.M, T.K, M.M.

Contributed to interpretation of results: L.P, D.J., E.J, A.M, L.S, M.K, K.T, C.B.M, Le.S, G.C, M.M., P.G, M.L, A.B.M, M.P, E.M, M.G.T, A.Ma, R.N, R.V, E.W, T.K, M.M.

Wrote manuscript: L.P, D.J., E.J, A.M, F.C, G.H, M.D, A.E, A.C, M.A.W.S, B.Y, J.L, S.Ti, M.M., P.G, M.L, A.B.M, M.P, M.G.T, A.Ma, R.N, R.V, E.W, T.K, M.M.

Acknowledgements

Support was provided by: Estonian Research Infrastructure Roadmap grant no 3.2.0304.11-0312; Australian Research Council Discovery grants (DP110102635 and DP140101405) (D.M.L, M.W. and E.W.); Danish National Research Foundation; the Lundbeck Foundation and KU2016 (E.W.); ERC Starting Investigator grant (FP7 - 261213) (T.K.); Estonian Research Council grant PUT766 (G.C.; M.K.); EU European Regional Development Fund through the Centre of Excellence in Genomics to Estonian Biocentre; Estonian Institutional Research grant IUT24-1; (L.S.; M.J.; A.K.; B.Y.; K.T.; C.B.M.; Le.S.; H.Sa.; S.L.; D.M.B.; E.M.; R.V.; G.H.; M.K.; G.C.; T.K.; M.M.); French Ministry of Foreign and European Affairs and French ANR grant number ANR-14-CE31-0013-01 (F.-X.R.); Gates Cambridge Trust Funding (E.J.); ICG SB RAS (No. VI.58.1.1) (D.V.L.); Leverhulme Programme grant no. RP2011-R-045 (A.B.M., P.G. & M.G.T.); Ministry of Education and Science of Russia; Project 6.656.2014/K (S.A.F.); NEFREX grant funded by the European Union (People Marie Curie Actions; International Research Staff Exchange Scheme; call FP7-PEOPLE-2012-IRSES-number 318979) (M.M.; G.H.); NIH grants 5DP1ES022577 05, 1R01DK104339-01, and 1R01GM113657-01 (S.Tis); Russian Foundation for Basic Research (grant N 14-06-00180a) (M.G.); Russian

Foundation for Basic Research; grant 16-04-00890 (O.B.; E.B); Russian Science Foundation grant 14-14-00827 (O.B.); The Russian Foundation for Basic Research (14-04-00725-a), The Russian Humanitarian Scientific Foundation (13-11-02014) and the Program of the Basic Research of the RAS Presidium "Biological diversity" (E.K.K.); Wellcome Trust and Royal Society grant WT104125AIA & the Bristol Advanced Computing Research Centre - http://www.bris.ac.uk/acrc/ (D.J.L); Wellcome Trust grant 098051 (Q.A.; C.T.-S.; Y.X.); Wellcome Trust Senior Research Fellowship grant 100719/Z/12/Z (M.G.T); Young Explorers Grant from the National Geographic Society (8900-11) (C.A.E.); ERC Consolidator Grant 647787 'LocalAdaptatio' (AM); Program of the RAS Presidium "Basic research for the development of the Russian Arctic" (B.M.); Russian Foundation for Basic Research grant 16-06-00303 (E.B).

The authors declare no competing financial interests.

Figure and Table Legends

Table 1 Positive selection hits. A subset of novel positive selection findings in our 12 macro-regional groups defined using fineSTRUCTURE.

Figure 1 Genetic barriers across space. Spatial visualisation of genetic barriers inferred from genome-wide genetic distances, quantified as the magnitude of the gradient of spatially interpolated allele frequencies (value denoted by colour bar; grey areas have been land during the last glacial maximum but are currently under water). Here we used a novel spatial kernel smoothing method based on the matrix of pairwise average heterozygosity. **Inset:** partial correlation between magnitude of genetic gradients and combinations of different geographic factors, elevation (E), temperature (T) and precipitation (P), for genetic gradients from fineSTRUCTURE (red) and allele frequencies (blue). This analysis (SI1:2.2.2 for details) shows that despite the large number of prehistoric movements across Eurasia, genetic differences within this region have been strongly shaped by physical barriers such as mountain ranges, deserts, forests, and open water (such as the Wallace line).

Figure 2 Evidence of an xOoA component in the genome of modern Papuans. Panel A: MSMC split times plot. The Yoruba-Eurasia split curve shows the mean of all Eurasian genomes against one Yoruba genome. The grey area represents top and bottom 5% of runs. We chose a Koinanbe genome as representative of the Sahul populations. Panels B-D: Decomposition of the ChromoPainter inferred African chunks in Papuans. Panel **B**: Semi-parametric decomposition of the joint distribution of chunk lengths and non-African derived allele rate per SNP, showing the relative proportion of chunks in K=20 components of the distribution, ordered by non-African derived allele rate, relative to the overall proportion of chunks in each component. The four datasets produced by considering (African/Denisova) chunks in (Europeans/Papuans) are shown with our inferred "extra Out-of-Africa xOoA" component. Panel **C**: The reconstruction of African chunks in Papuans using a mixture of all other data (red) and with the addition of the xOoA component (black). Panel **D**: The properties of the components in terms of non-African derived allele rate, on which the components are ordered, and length.

Extended Display Items

ED1 Sample Diversity and Archaic signals. **A**: Map of location of samples highlighting the Diversity/Selection Sets; **B**: ADMIXTURE plot (K=8 and 14) which relates general visual inspection of genetic structure to studied populations and their region of origin; **C**: Sample level heterozygosity is plotted against distance from Addis Ababa. The trend line represents only non-African samples. The inset shows the waypoints used to arrive at the distance in kilometres for each sample. **D**: Boxplots were used to visualize the Denisova (red), Altai (green) and Croatian Neanderthal (blue) D distribution for each regional group of samples. Oceanian Altai D values show a remarkable similarity with the Denisova D values for the same region, in contrast with the other groups of samples where the Altai boxplots tend to be more similar to the Croatian Neanderthal ones.

ED2 Data quality checks and heterozygosity patterns. Concordance of DNA sequencing (Complete Genomics Inc.) and DNA genotyping (Illumina genotyping arrays) data (ref-ref; het-ref-alt and hom-alt-alt, see SI 1.6) from chip (**A**) and sequence data (**B**). Coverage (depth) distribution of variable positions, divided by DNA source (Blood or Saliva) and Complete Genomic calling pipeline (release version) (**C**). Genome-wide distribution of Transition/Transversion ratio subdivided by DNA source (Saliva or Blood) and by Complete Genomic calling pipeline (**D**). Genome-wide distribution of Transition/Transversion ratio subdivided by chromosomes (**E**). Inter-chromosome differences in observed heterozygosity in 447 samples from the Diversity Set (**F**). Inter-chromosome differences in observed heterozygosity in a set of 50 unpublished genomes from the Estonian Genome Center, sequenced on an Illumina platform at an average coverage exceeding 30x (**G**). Inter-chromosome differences in observed heterozygosity in the phase 3 of the 1000 Genomes Project (**H**). The total number of observed heterozygous sites was divided by the number of accessible basepairs reported by the 1000 Genomes Project.

ED3: FineSTRUCTURE Chunk counts or 'co-ancestry'. ChromoPainter and FineSTRUCTURE results, showing both inferred populations with the underlying (averaged) number of "haplotype chunks" that an individual in a population receives (rows) from donor individuals in other populations (columns). 108 populations are inferred by FineSTRUCTURE. The dendrogram shows the inferred relationship between populations. The numbers on the dendrogram give the proportion of MCMC iterations for which each population split is observed (where this is less than 1). Each "geographical region" has a unique colour from which individuals are labeled. The number of individuals in each population is given in the label; e.g. "4Italians; 3Albanians" is a population of size 7 containing 4 individuals from Italy and 3 from Albania.

ED4: MSMC genetic split times and outgroup f3 results. The MSMC split times estimated between each sample and a reference panel of 9 genomes were linearly interpolated to infer the broader square matrix (**A**). Summary of outgroup f3 statistics for each pair of non-African populations (**B**) or to an ancient sample (**C**) using Yoruba as an outgroup. Populations are grouped by

geographic region and are ordered with increasing distance from Africa (left to right for columns and bottom to top for rows). Colour bars at the left and top of the heat map indicate the colour coding used for the geographical region. Individual population labels are indicated at the right and bottom of the heat map. The f3 statistics are scaled to lie between 0 and 1, with a black colour indicating those close to 0 and a red colour indicating those close to 1. Let m and M be the minimum and maximum f3 values within a given row (i.e., focal population). That is, for focal population X (on rows), m = minY,Y≠X f3(X, Y ; Yoruba) and M = maxY,Y≠X f3(X, Y ; Yoruba). The scaled f3 statistic for a given cell in that row is given by f3scaled=(f3-m)/(M-m), so that the smallest f3 in the row has value f3scaled=0 (black) and the largest has value f3scaled=1 (red). By default, the diagonal has value f3scaled=1 (red). The heat map is therefore asymmetric, with the population closest to the focal population at a given row having value f3scaled=1 (red colour) and the population farthest from the focal population at a given row having value f3scaled=0 (black colour). Therefore, at a given row, scanning the columns of the heat map reveals the populations with the most shared ancestry with the focal population of that row in the heat map.

ED5: **Geographical patterns of genetic diversity.** Isolation by distance pattern across areas of high genetic gradient, using Europe as a baseline. The samples used in each analysis are indicated by coloured lines on the maps to the right of each plot. The panels show F_{ST} as a function of distance across the Himalayas (**A**), the Ural mountains (**B**), and the Caucasus (**C**) as reported on the color-coded map (**D**). Effect of creating gaps in the samples in Europe (**E**): we tested the effect of removing samples from stripes, either north to south (**F**) or west to east (**G**), to create gaps comparable in size to the gaps in samples in the dataset. Effective migration surfaces inferred by EEMS (**H**).

ED6: Summary of positive selection results: Barplot comparing frequency distributions of functional variants in Africans and non-Africans (**A**). The distribution of exonic SNPs according to their functional impact (synonymous, missense and nonsense) as a function of allele frequency. Note that the data from both groups was normalised for a sample size of n=21 and that the Africans show significantly ($P<10^{-15}$) more rare variants across all sites classes.
Result (**B**) of 1000 bootstrap replica of the Rxy test for a subset of pigmentation genes highlighted by GWAS (n=32). The horizontal line provides the African reference (x=1) against which all other groups are compared. The blue and red marks show the 95^{th} and the 5^{th} percentile of the bootstrap distributions respectively. If the 95^{th} percentile is below 1, then the population shows a significant excess of missense variants in the pigmentation subset relative to the Africans. Note that this is the case all non-Africans except the Oceanians. Pools (**C**) of individuals for selection scans. fineSTRUCTURE based coancestry matrix was used to define twelve groups of populations for the downstream selection scans. These groups are highlighted in the plot by boxes with broken line edges. The number of individuals in each group is reported in Table SI2:3.2-I.

ED7: African chunk lengths as a function of genome proportion, for different painting scenarios. A: 447 Diversity Panel results, showing label averages (large crosses) along with individuals (small dots). **B**: Relative excluded

Diversity Panel results, to check for whether including related individuals affects African genome fraction. Individuals that shared more than 2% of genome fraction were forbidden from receiving chunks from each other, and the painting was re-run on a large subset of the genome (all ROH regions from any individual). **C**: ROH only African chunks. To guard against phasing errors, we painted only regions for which an individual was in a long (>500kb) Run of Homozygosity using the PLINK command "--homozyg-window-kb 500000 --homozyg-window-het 0 --homozyg-density 10". Because there are so few such regions, we report only the population average for populations with two or more individuals, as well as the standard error in that estimate. Populations for whom the 95% CI passed 0 were also excluded. Note the logarithmic axis. **D**: Ancient DNA panel results. We used a different panel of 109 individuals which included 3 ancient genomes. We painted Chromosomes 11, 21 & 22 and report as crosses the population averages for populations with 2 or more individuals. The solid thin lines represent the position of each population when modern samples only are analysed. The dashed lines lead off the figure to the position of the ancient hominins and the African samples.

ED8: MSMC Linear behavior of MSMC split estimates in presence of admixture. The examined Central Asian (**A**), East African (**B**), and African-American (**C**) genomes yielded a signature of MSMC split time (Truth, left-most column) that could be recapitulated (Reconstruction, second left most column) as a linear mixture of other MSMC split times. The admixture proportions inferred by our method (top of each admixture component column) were remarkably similar to the ones previously reported from the literature.
MSMC split times (**D**) calculated after re-phasing an Estonian and a Papuan (Koinanbe) genome together with all the available West African and Pygmy genomes from our dataset to minimize putative phasing artefacts. The cross coalescence rate curves reported here are quantitatively comparable with the ones of Figure 2 A, hence showing that phasing artefacts are unlikely to explain the observed past-ward shift of the Papuan-African split time. Boxplot (**E**) showing the distribution of differences between African-Papuan and African-Eurasian split times obtained from coalescent simulations assembled through random replacement to make 2000 sets of 6 individuals (to match the 6 Papuans available from our empirical dataset), each made of 1.5 Gb of sequence. The simulation command line used to generate each chromosome made of 5Mb was as follows, being *DIV*=0.064; 0.4 or 0.8 for the xOoA, Denisova (Den) and Divergent Denisova (DeepDen) cases, respectively: ms0ancient2 10 1 .065 .05 -t 5000. -r 3000. 5000000 -I 7 1 1 1 1 2 2 2 -en 0. 1 .2 -en 0. 2 .2 -en 0. 3 .2 -en 0. 4 .2 -es .025 7 .96 -en .025 8 .2 -ej .03 7 6 -ej .04 6 5 -ej .060 8 3 -ej .061 4 3 -ej .062 2 1 -ej .063 3 1 -ej *DIV* 1 5

ED9: Modelling the xOoA components with FineSTRUCTURE. A: Joint distribution of Chunk lengths and Derived allele count, showing the median position of each cluster and all chunks assigned to it in the Maximum A Posteriori (MAP) estimate. Note that although a different proportion of points is assigned to each in the MAP, the total posterior is very close to 1/K for all. The dashed lines show a constant mutation rate. Chunks are ordered by mutation rate from low to high. **B**: Residual distribution comparison between the two

component mixture using EUR.AFR and EUR.PNG (left), and the three component mixture including xOoA (using the same colour scale) (right). The residuals without xOoA are larger (RMSE 0.0055 compared to RMSE 0.0018) but more importantly, they are also structured. **C**: Assuming a mutational clock and a correct labeling of chunks, we can estimate the relative age of the splits from the number of derived alleles observed on the chunks. This leads to an estimate of 1.5 times older for xOoA compared to the Eurasian-Africa split.

ED10: Cartoon illustration of the proposed xOoA model. A subway map figure illustrating, as suggested by the novel results presented here, the model of an early, extinct Out-of-Africa (xOoA) entering the genome of Sahul populations at their arrival in the region. Given the overall small genomic contribution of this event to the genome of modern Sahul, we could not determine whether the documented Denisova admixture (question marks) and putative multiple Neanderthal admixtures took place along this extinct OoA. We also speculate (question mark) people who migrated along the xOoA route may have left a trace in the genome of the Altai Neanderthal as reported by Kuhlwilm and colleagues[12].

Methods

Data Preparation: We analyse a set of genomes sequenced by the same technology (Complete Genomics Inc.) which results in minimal platform differences between batches of samples analysed by slight modifications of CG proprietary pipeline (ED2; SI 1.6). We see good concordance between CG sequence and Illumina genotyping array results for the same samples with minor reference bias in the latter data (ED2; SI 1.6). In the final dataset, we retained only one second (Australians, to make use of all the available samples) - and five third-degree relatives pairs (SI2:1.7-I). All genomes were annotated against the Ensembl GRCh37 database and compared to dbSNP Human Build 141 and Phase 1 of the 1000 Genomes Project dataset[29] (SI1:1.1-6). We found 10,212,117 new SNPs, 401,911 of which were exonic. As expected from our sampling scheme, existing lists of variable sites have been extended mostly by the Siberian, South-East Asian and South Asian genomes, which contribute 89,836 (22.4%), 63,964 (15.9%) and 40,758 (10.1%) of the new exonic variants detected in this study.
Compared to the genome-wide average, we see fewer heterozygous sites on chromosomes 1 and 2, and an excess on chromosomes 16, 19 and 21 (ED2). This pattern is independent of simple potential confounders, such as rough estimates of recombination activity and gene density (SI1:1.8), and mirrors the inter-chromosomal differences in divergence from chimpanzee[31], suggesting large-scale differences in mutation rates among chromosomes. We confirmed this general pattern using 1000Genomes Project data (SI1:1.8).
The "ancient genome diversity panel" consisted of 106 samples from the main Diversity panel along with Altai Neanderthal, Denisova and the Modern Human reference genome. Sites that are heterozygous in archaic humans were removed.

Geographic gradient analyses. We used a Gaussian kernel smoothing (based on the shortest distance on land to each sample) to interpolate genetic patterns across space. Averaging over all markers, we obtained an expression for the mean square gradient of allele frequencies in terms of the matrix of genetic distance between pairs of samples (SI1:2.2.2). This provides a simple way to identify spatial regions that contribute strongly to genetic differences between samples, and can be used, in principle, for any measure of genetic difference (for

641 fineSTRUCTURE data, we used negative shared haplotype length as a measure of
642 differentiation).
643
644 To quantify the link between the magnitude of genetic gradients (from
645 fineSTRUCTURE and allele frequency data) and geographic factors, we fitted a
646 generalised linear model to the sum of genetic magnitude gradients on the
647 shortest paths between samples to elevation, minimum quarterly temperature,
648 and annual precipitation summed in the same way, controlling for path length
649 and spatial random effects (SI1:2.2.2), and calculated partial correlations
650 between genetic gradient magnitudes and geographic factors.
651
652 **Finestructure Analysis.** FineSTRUCTURE[32] was run as described in
653 SI1:S2.2.1. Within the 106 genetically distinct genetic groups, labels were
654 typically genetically homogeneous - 113 of the 148 population labels (76%)
655 were assigned to only one 'genetic cluster'. Similarly, genetic clusters were
656 typically specific to a label, with 66 of the 106 'genetic clusters' (62%) containing
657 only one population label.
658 *Correction for phasing errors:* To check whether phasing errors could produce
659 the shorter Papuan chunks, we focussed on regions of the genome that had an
660 extended (>500Kb) run of homozygosity. We ran ChromoPainter for each
661 individual on only these regions, meaning each individual was only painted
662 where it had been perfectly phased. This did not change the qualitative features
663 (SI1:2.2.1).
664 *Removal of similar samples:* Papuans are genetically distinct from other
665 populations due to tens of thousands of years of isolation. We wanted to check
666 whether African chunk lengths were biased by the inclusion of a large number of
667 relatively homogeneous Eurasians with few Papuans. To do this we repeated the
668 N=447 painting allowing only donors from dissimilar populations, including only
669 individuals who donated <2% of a genome in the main painting. This did not
670 change the qualitative chunk length features (SI1:2.2.1).
671 *Inclusion of ancient samples:* We ran our smaller individual panel with (N=109)
672 and without (N=106) ancient samples (Denisova, Neanderthal and ancestral
673 human). This did not change the qualitative chunk length features (SI1:2.2.1).

Selection Analyses. We investigated balancing, positive and purifying selection for a part of the dataset with larger group sizes which was defined as the Selection subset (SI2:3.1-I; SI2:3.2-I) using a wide range of window-based as well as variant-based approaches. Furthermore we investigated how these signals relate to shared demographic history. Where possible we contextualized our findings by integrating them with information from various functional databases. Detailed descriptions of all methods used are available in SI1:3.

MSMC, Denisova masking, simulations of alternative scenarios and assessment of phasing robustness. Genetic split times were initially calculated following the standard MSMC procedure[8], and subsequently modified as follows. To estimate the effect of archaic admixture, putative Denisova haplotypes were identified in Papuans using a previously published method[27] and masked from all the analysed genomes. Particularly, whether a putative archaic haplotype was found in heterozygous or homozygous state within the chosen Papuan genome, the "affected" locus was inserted into the MSMC mask files and, hence, removed from the analysis.

We note that a fraction of the Denisova and Neanderthal contributions to the Papuan genomes may be indistinguishable, due to the shared evolutionary history of these two archaic populations. As a result, some of the removed "Denisova" haplotypes may have actually entered the genome of Papuans through Neanderthal. Regardless of this, our exercise successfully shows that the MSMC split time estimates are not affected by the documented presence of archaic genomic component (whether coming entirely from Denisova or partially shared with Neanderthal).

We further excluded the role of Denisova admixture in explaining the deeper African-Papuan MSMC split times through coalescent simulations (using ms to generate 30 chromosomes of 5 Mbp each, and simulating each scenario 30 times). These showed that the addition of 4% Denisova lineages to the Papuan genomes does not change the MSMC results, while the addition of 4% xOoA lineages recreates the qualitative shift observed in the empirical data.

Phasing artefacts were also taken into account as putative confounders of the MSMC split time estimates. We re-run MSMC after re-phasing one Estonian, one Papuan and 20 West African and Pygmies genomes in a single experiment. By this way we ruled out potential artefacts stemming from the excess of Eurasian over Sahul samples during the phasing process. Both the archaic and phasing corrections yielded the same split time as of the standard MSMC runs.

Emulation of all pairwise MSMC split times. We confirmed that none of the other populations behaved as an outlier from those identified in the N=22 full pairwise analysis by estimating the MSMC split times between all pairs. We chose 9 representative populations (including Papuan, Yoruba and Baka) from the 22, and compared each of the 447 diversity panel genomes to them. We learn a model for each individual l not in our panel,

$\hat{t}_{lj} = \sum_{k=1}^{9} \alpha_{lk} t_{lj}$ for $j \in (1..9)$,

where the positive mixture weights α_k sum to 1 and are otherwise learned from the $j \in (1..9)$ observations which we have data under quadratic loss. We can then predict the unobserved values

$$\hat{t}_{li} = \sum_{k=1}^{9} \alpha_k t_{ki}.$$

Examination of this matrix (SI1:S2.2.3, SI2:2.2.3-III) implies no other populations are expected to have unusual MSMC split times from Africa.

Mixture model for African haplotypes in Papuans. *Obtaining haplotypes from painting:* We define as African or Archaic chunk in Eurasians or Papuans a genomic locus spanning at least 1000bp, and showing SNPs that were assigned by chromopainter a 50% chance of copying from either an African or Archaic genome, respectively. For each chunk we then calculated the number of non-African mutations, defined as sites found in derived state in a given chunk and in ancestral state in all of the African genomes included in the present study. *Modelling:* We used a non-parametric model for the joint distribution of length and non-African derived allele mutation rate of chunks. We fit K (=20) components to the joint distribution. Each component has a characteristic length

l_k, variability σ_k and mutation rate μ_k. A chunk of length l_i with X_i such mutations from component $I_i = k$ has the following distribution:

$l_i|\{l_k,\sigma_k^2,I_i = k\}$~log-Normal$(l_k,\sigma_k^2)$

$X_i|\{l_k,\mu_k,I_i = k\}$~Binomial$(l_k,\mu_k)$

This model for chunk lengths is motivated by the extreme age of the split times we seek to model. Recent splits would lead to an exponential distribution of haplotype lengths. However, due to haplotype fixation caused by finite population size, very old splits have finite (non-zero) haplotype lengths. Additionally, the data are left-censored since we cannot reliably detect chunks that are very short. We note that whilst this makes a single component a reasonable fit to the data, as K increases the specific choice becomes less important.

We then impose the prior $p(I_i = k) = 1/K$ and use the Expectation-Maximization algorithm to estimate the mixture proportions $\pi_{ik} = E(I_{ik}|l_i,X_i)$ along with the maximum likelihood parameter estimates $\{l_k,\sigma_k^2,\mu_k\}$. We do this for the four combinations of African (AFR) and Denisova (DEN) chunks found in Papuans (PNG) or Europeans (EUR), in order to learn the parameters. SI1:S2.2.10 describes this in more detail. We then describe the distribution of chunks for each class c of chunk in terms of the expected proportion of chunks found in each component,

$\pi_{ck} = \frac{\pi_{ck}^{'}}{\sum_{k=1}^{K}\pi_{ck}^{'}}$ where $\pi_{ck}^{'} = \sum_{i=1}^{N_c}\pi_{cik}$,

where N_c is the number of chunks of class c. π_c is a vector of the proportions from each of the K components.

<u>*Single-out-of-Africa model:*</u> We fit African chunks in Papuans as a mixture of the others in a second layer of mixture modelling:

$$\pi_{PNG.AFR} = \sum_{c \in \{PNG.DEN,EUR.AFR,EUR.DEN\}} \alpha_c\,\pi_c,$$

where α_c sum to 1. This is straightforward to fit.

<u>*xOoA model:*</u> We jointly estimate an additional component π_{xOoA} and the mixture contributions β_c under the mixture

$$\pi_{PNG.AFR} = \sum_{c \in \{PNG.DEN, EUR.AFR, EUR.DEN, xOoA)} \beta_c \, \pi_c.$$

This is non-trivial to fit. We use a penalisation scheme to simultaneously ensure we a) obtain a valid mixture for β_c, b) give a prediction x_k that is also a valid mixture, c) leave little signal in the residuals, and d) obtain a good fit. Cross-validation is used to obtain the optimal penalisation parameters (A and B) with the loss function:

$$\text{loss} = \sum_{k=1}^{K} e_k^2 + AP_A + BP_B,$$

where e_k are the residuals in each component, $P_A = |(\sum_c \beta_c) - 1| + |(\sum_k x_k) - 1|$ (for a valid mixture) and $P_B = s.d(e_k)$ (for requirement c, good solutions will have similar residuals across components). The loss is minimised via standard optimization techniques. SI1:S2.2.10 details how initial values are found and explores the robustness of the solution to changes in A and B - the results do not change qualitatively for reasonable choices of these parameters, and the mixtures are valid to within numerical error.

Genome-wide xOoA estimation: We used the estimated xOoA derived allele mutation rate estimate θ_{xOoA} to estimate the xOoA contribution in haplotypes classed as Eurasian or Papuan by ChromoPainter. First we obtained estimates of $\pi_{PNG.EUR}$ and $\pi_{PNG.PNG}$ using the single out-of-Africa model above, additionally allowing a EUR.EUR contribution. We then estimate α_{xOoA} using the observed mutation rate θ_{obs} and that predicted under the mixture model θ_{mix} by rearranging the mixture:

$$\theta_{obs} = \alpha_{xOoA}\theta_{xOoA} + (1 - \alpha_{xOoA})\theta_{mix}$$

Estimates less than zero are set to 0. The genome wide estimate is obtained by weighting each θ by the proportion of the genome that was painted with that donor. Neanderthal and Denisova chunks were assumed to be proxied by PNG.DEN (0% xOoA by assumption); African chunks by PNG.AFR; Papuan and Australian by PNG.PNG and all other chunks by PNG.EUR. We obtain confidence intervals by bootstrap resampling of haplotypes for each donor/recipient pair.

We estimate the proportion of xOoA in Papuan chunks assigned as both Eurasian (0.1%, 95% CI 0-2.6) and Papuan (4%, 95% CI 2.9-4.5) (SI1:2.2.10), by using the estimated mutation density in xOoA.

Y chromosome and mtDNA haplopgroup analysis. The presence of an extinct xOoA trace in the genome of modern Papuans may seem at odds with analyses of mtDNA and Y chromosome phylogenies, which point to a single, recent origin for all non-African lineages (mtDNA L3, which gives rise to all mtDNA lineages outside Africa has been dated at ~70 kya,[33,34]). However, uniparental markers inform on a small fraction of our genetic history, and a single origin for all non-African lineages does not exclude multiple waves OoA from a shared common ancestor. We show analytically (SI1:2.2.12) that, if the xOoA entered the Papuan genome >40 kya, their mtDNA and Y lineages could have been lost by genetic drift even assuming an initial xOoA mixing component of up to 35%. Similar findings have been reported recently[13].

References

1 Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78-81, doi:10.1126/science.1181498 (2010).

2 Lachance, J. *et al.* Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. *Cell* **150**, 457-469, doi:10.1016/j.cell.2012.07.009 (2012).

3 Pagani, L. *et al.* Tracing the Route of Modern Humans out of Africa by Using 225 Human Genome Sequences from Ethiopians and Egyptians. *American journal of human genetics* **96**, 986-991, doi:10.1016/j.ajhg.2015.04.019 (2015).

4 Clemente, F. J. *et al.* A Selective Sweep on a Deleterious Mutation in CPT1A in Arctic Populations. *American journal of human genetics* **95**, 584-589, doi:10.1016/j.ajhg.2014.09.016 (2014).

5 Gudbjartsson, D. F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nat Genet* **47**, 435-444, doi:10.1038/ng.3247 (2015).

6 Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* **45**, 1113-1120, doi:10.1038/ng.2764 (2013).

7 Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493-496, doi:10.1038/nature10231 (2011).

8 Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat Genet* **46**, 919-925, doi:10.1038/ng.3015 (2014).

9 Patterson, N. *et al.* Ancient admixture in human history. *Genetics* **192**, 1065-1093, doi:10.1534/genetics.112.145037 (2012).

10 Scally, A. & Durbin, R. Revising the human mutation rate: implications for understanding human evolution. *Nat Rev Genet* **13**, 745-753, doi:10.1038/nrg3295 (2012).

11 Grove, M. *et al.* Climatic variability, plasticity, and dispersal: A case study from Lake Tana, Ethiopia. *Journal of human evolution* **87**, 32-47, doi:10.1016/j.jhevol.2015.07.007 (2015).

12 Kuhlwilm, M. *et al.* Ancient gene flow from early modern humans into Eastern Neanderthals. *Nature* **530**, 429-433, doi:10.1038/nature16544 (2016).

13 Groucutt, H. S. *et al.* Rethinking the dispersal of Homo sapiens out of Africa. *Evol Anthropol* **24**, 149-164, doi:10.1002/evan.21455 (2015).

14 Liu, W. *et al.* The earliest unequivocally modern humans in southern China. *Nature* **526**, 696-699, doi:10.1038/nature15696 (2015).

15 Reyes-Centeno, H. *et al.* Genomic and cranial phenotype data support multiple modern human dispersals from Africa and a southern route into Asia. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 7248-7253, doi:Doi 10.1073/Pnas.1323666111 (2014).

16 Mellars, P., Gori, K. C., Carr, M., Soares, P. A. & Richards, M. B. Genetic and archaeological perspectives on the initial modern human colonization of southern Asia. *Proceedings of the National Academy of Sciences of the United*

States of America **110**, 10699-10704, doi:Doi 10.1073/Pnas.1306043110 (2013).
17 Prugnolle, F., Manica, A. & Balloux, F. Geography predicts neutral genetic diversity of human populations. *Current Biology* **15**, R159-R160 (2005).
18 Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710-722, doi:10.1126/science.1188021 (2010).
19 Reich, D. *et al.* Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. *American journal of human genetics* **89**, 516-528, doi:10.1016/j.ajhg.2011.09.005 (2011).
20 Fu, Q. *et al.* Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* **514**, 445-449, doi:10.1038/nature13810 (2014).
21 Fu, Q. *et al.* A revised timescale for human evolution based on ancient mitochondrial genomes. *Current Biology* **23**, 553-559, doi:10.1016/j.cub.2013.02.044 (2013).
22 Fu, Q. *et al.* The genetic history of Ice Age Europe. *Nature*, doi:10.1038/nature17993 (2016).
23 Meyer, M. *et al.* A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science* **338**, 222-226, doi:Doi 10.1126/Science.1224344 (2012).
24 Petkova, D., Novembre, J. & Stephens, M. Visualizing spatial population structure with estimated effective migration surfaces. *Nat Genet* **48**, 94-100, doi:10.1038/ng.3464 (2016).
25 Hellenthal, G. *et al.* A genetic atlas of human admixture history. *Science* **343**, 747-751, doi:10.1126/science.1243518 (2014).
26 Chapman, N. H. & Thompson, E. A. A model for the length of tracts of identity by descent in finite random mating populations. *Theoretical population biology* **64**, 141-150 (2003).
27 Wall, J. D. *et al.* Higher levels of neanderthal ancestry in East Asians than in Europeans. *Genetics* **194**, 199-209, doi:10.1534/genetics.112.148213 (2013).
28 Posth, C. *et al.* Pleistocene Mitochondrial Genomes Suggest a Single Major Dispersal of Non-Africans and a Late Glacial Population Turnover in Europe. *Current biology : CB*, doi:10.1016/j.cub.2016.01.037 (2016).
29 The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65, doi:10.1038/nature11632 (2012).
30 Migliano, A. B., Vinicius, L. & Lahr, M. M. Life history trade-offs explain the evolution of human pygmies. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 20216-20219, doi:10.1073/pnas.0708024105 (2007).

These references only appear in the online Methods

31 Mikkelsen, T. *et al.* Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69-87 (2005).
32 Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. *PLoS genetics* **8**, e1002453, doi:10.1371/journal.pgen.1002453 (2012).

33 Behar, D. M. *et al.* A "Copernican" Reassessment of the Human Mitochondrial DNA Tree from its Root. *American journal of human genetics* **90**, 675-684, doi:Doi 10.1016/J.Ajhg.2012.03.002 (2012).

34 Soares, P. *et al.* The Archaeogenetics of Europe. *Current Biology* **20**, R174-R183 (2010).

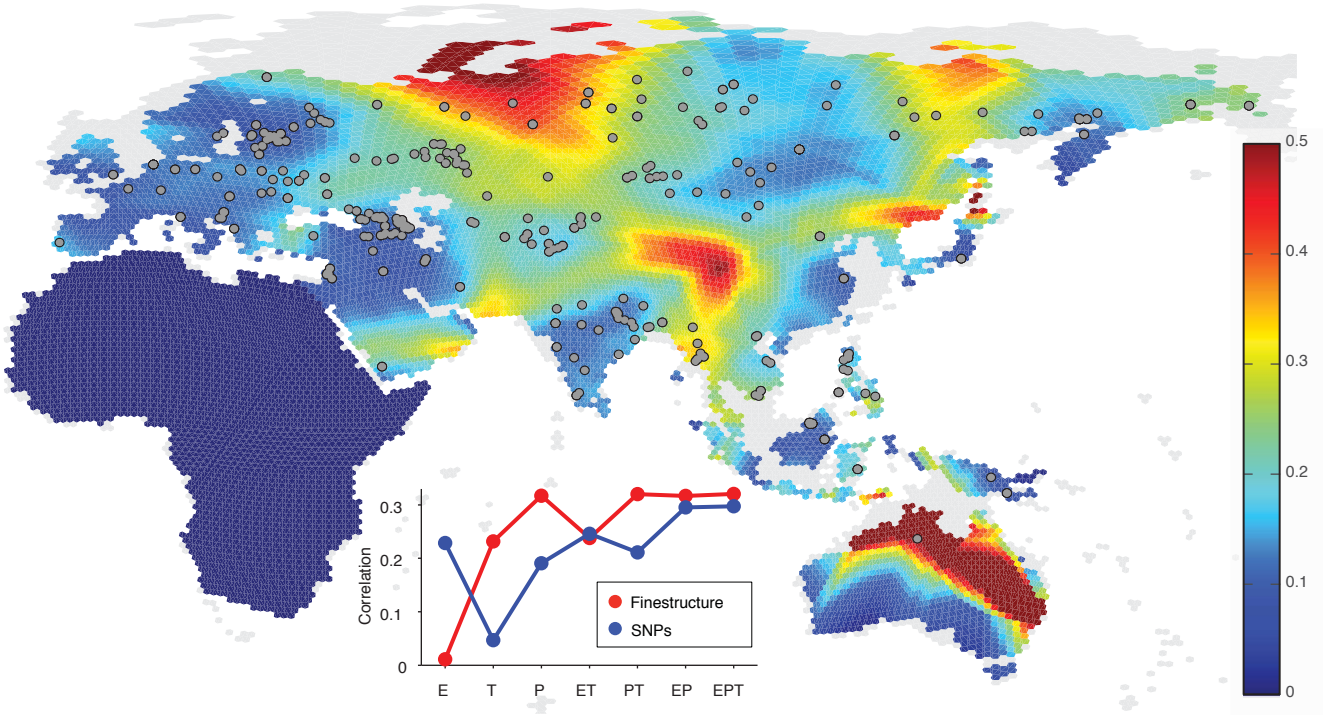

0.5
0.4
0.3
0.2
0.1
0
Correlation
0.3
0.2
0.1
0
E
T
P
ET
PT
EP
EPT
Finestructure
SNPs

Table 1 Eurasian subset of variants highlighted by positive selection tests

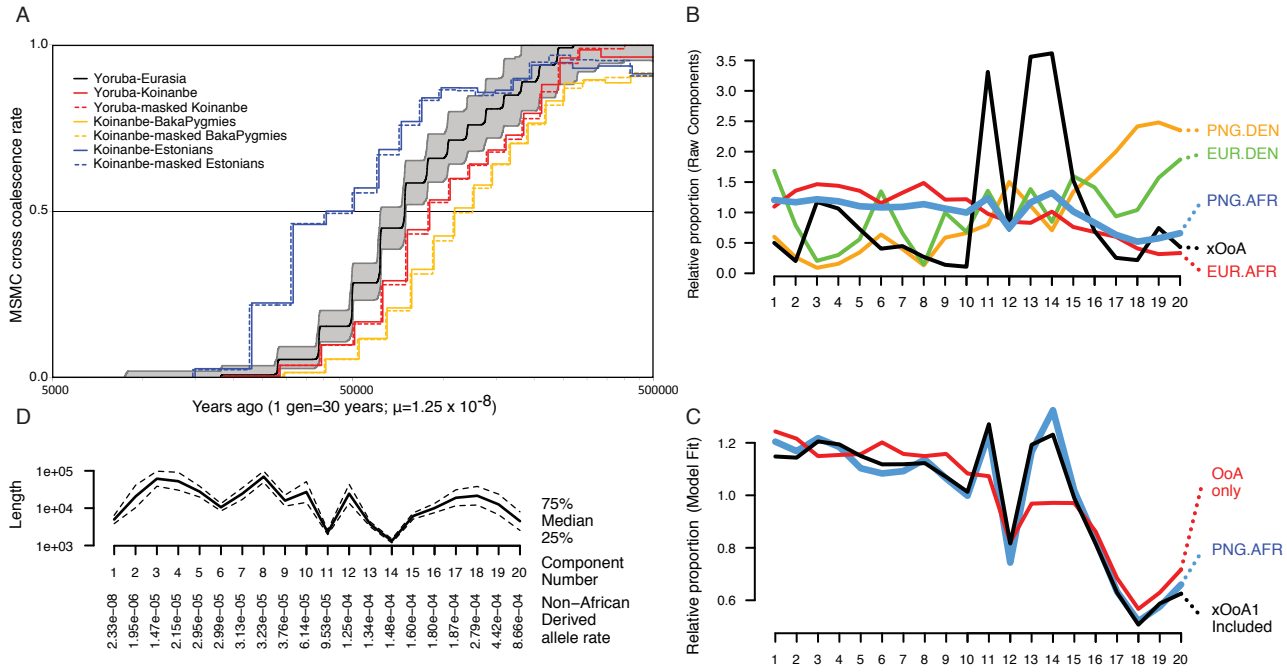
Gene	SNP	Variant Type	Test	Population	Phenotype
FADS2	rs2524296	intronic	di	Wsi	Fatty acid desaturation
ZNF646	rs749670	missense	dDAF,DIND	CSi	Lipid metabolism, bile synthesis
PPARA	rs6008197	missense	iHS,nSL,TD,DIND	SoA	Lipid metabolism
GANC	rs8024732	missense	iHS,DIND	SoA	Carbohydrate metabolism
PKDREJ	rs6519993	missense	iHS,nSL,TD,DIND	SoA	Sperm-Receptor, kidney disease
CSMD1	rs7816731	non-coding	di	Wsi	Blood pressure
LYPD3	rs117823872	non-coding	di	Wsi	Wound healing
POU2F3	rs882856	missense	dDAF	WEu	Wound healing
B9D1	rs4924987	missense	dDAF	EEu	Ciliogenesis
PCDH15	rs4935502	missense	dDAF	CSi	Ciliogenesis
TMEM216	rs10897158	missense	dDAF	Wsi	Ciliogenesis
PLCB2	rs936212	missense	dDAF	NSi	Ciliogenesis
MYO18B	rs2236005	missense	dDAF	SeI	Motor activity
FLNB	rs12632456	missense	dDAF	SeI	Motor activity
TTN	rs10497520	missense	dDAF	MiE	Motor activity

Note the abbreviations of the population group names are according to Table S2.2

iHS,nSL, or TD, indicates that the variant is a from a top 1% window by that test for the indicated population. DIND indicates that the variant is significantly (>5SD) above the neutral background by the DIND test (See Supplementary Section 3)

di indicates that the variant was in the top 12 of the most highly divergent SNVs by the di score in each of the twelve population groups (See Supplementary Section 3)

dDAF indicates that the variant was in the top 20 most highly differentiated SNPs in its class in a given comparison (See Supplementary Section 3)


A
B
C
D
MSMC cross coalescence rate
Yoruba-Eurasia
Yoruba-Koinanbe
Yoruba-masked Koinanbe
Koinanbe-BakaPygmies
Koinanbe-masked BakaPygmies
Koinanbe-Estonians
Koinanbe-masked Estonians
Years ago (1 gen=30 years; μ=1.25 x 10^-8)
Relative proportion (Raw Components)
PNG.DEN
EUR.DEN
PNG.AFR
xOoA
EUR.AFR
Relative proportion (Model Fit)
OoA only
PNG.AFR
xOoA1 Included
Length
75%
Median
25%
Component Number
Non−African Derived allele rate